2016

Statistical Cinema

Matthew Kehling
Xavier University - Cincinnati, kehlingm@xavier.edu

Statistical Cinema

Matthew Kehling
Xavier University

April 19, 2016

Abstract

Have you ever wondered what determines the length of a movie? What about the rating that movie gets on popular movie rating sites like Rottentomatoes.com or IMDb.com? In this paper, we explore a large movie data set and examine some relationships between the length of the movie and various attributes, such as: year released, genre, MPAA rating, and budget. We also apply various correlation measures (Pearson, Spearman, and Kendall, respectively) to audience ratings between two popular movie rating sites, Rottentomatoes.com and IMDb.com. In addition, we also describe the technique of web scraping for obtaining data from online sources, and we implement a bootstrapping approach to estimate the standard errors of the correlation statistics. So, please, sit back, relax, and enjoy the show...

1 Introduction

1.1 Data

The movie data set consists of 569 observations, or movies, with 23 variables, or different movie attributes; 17 of those were from IMDb.com, and the other 6 from Rottentomatoes.com. The variables are listed below:

- "title" : the name of the movie on IMDb.com
- "year" : the year in which the movie was released into theaters
- "length" : the length of the movie in minutes
- "mpaa" : the Motion Picture Association of America (MPAA) rating. This is the film's suitability for certain audiences based on the movie content, and has five levels: G, PG, PG-13, R, and Not Rated
- "rating" : the overall IMDb rating. This is the number enclosed in the gold star on all IMDb movie pages and can range from 0.0 to 10.0 with each rating holding one decimal place.

- "votes" : the number of votes used to determine the IMDb rating
- 9 genre indicator variables : in this data set there are nine genre variables that include Action, Adventure, Animation, Comedy, Drama, Documentary, Romance, Sci-fi, and Thriller. Each of these genres are variables in the data set with a "0" or a "1." A "1" indicates the corresponding movie is considered to have that genre, while a "0" indicated the corresponding movie does not have that genre. Current film theory has various reasons for assigning a genre to a film. (However, there is no indication on how these genres are assigned and put on IMDb. It is possible for a movie to have more than one genre, which is why the information has been collected this way. For example, the movie titled Big Hero 6 is considered an action, adventure, animation, comedy, and sci-fi movie and therefore has a "1" in each of the columns corresponding to those genres, and "0"s in the remaining genre columns.)
- "budget" : the approximate budget in US dollars of the movie
- "All Critic Score" : the average all critics' score on Rottentomatoes.com
- "All Critic Count" : the number of all critics used for the "All Critic Score"
- "Top Critic Score" : the average top critics' score on Rottentomatoes.com
- "Top Critic Count" : the number of top critics used for the "Top Critic Score"
- "Audience Score" : the average audience score on Rottentomatoes.com
- "Audience Count" : the number of audience voters used for the "Audience Score"

The last six variables listed above were obtained by a web scraping process described in [6]. The code developed for this paper can be found in Appendix A.

2 Initial Results

Once the data collection was complete, preliminary analysis was performed on the data. Outliers were observed and identified. The drastic outliers were found in "Audience Counts" variable, and the results were surprising; for example, the average Audience Score was 2,048,737.16 votes, while the median Audience Score was 255,326.50 votes. To illustrate how unusual this is, Figure 1 shows a graph of the IMDb Rating compared to Audience Score colored and sized according to Audience Count.

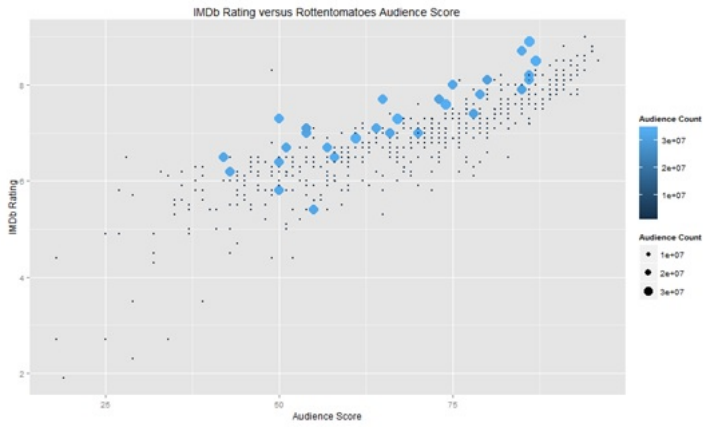


Figure 1: Audience Count Outliers

The 29 large blue dots compared are clearly outliers in this data set, as compared to the rest of the Audience Counts. Some of these movies include *Harry Potter and the Goblet of Fire*, *Wedding Crashers*, *Donnie Darko*, *Star Wars: Revenge of the Sith*, *Finding Nemo*, and *The Matrix*. At this time there is no explanation as to why 29 of the 569 movies have more than 30,000,000, when the next largest count under 30,000,000 is approximately 4,000,000, audience votes on Rottentomatoes. Another interesting phenomena, is that all of the 29 movies were movies release prior to 2005. I contacted Rottentomatoes Support for an explanation of these values, and, at the time of submitting this paper, have not heard back from them. Some possible explanations are some kind of reward for voting for these specific movies, a "fat finger" error when manually entering data into the database, or an error that occurred when transferring data to the Rottentomatoes database. However, for the remaining analysis, these movies were omitted.

Replotting the same graph without the outliers paints a more accurate picture of IMDb rating plotted against the Rottentomatoes Audience Score (See Figure 2). This graph illustrates the relationship between Audience Score on Rottentomatoes against the Rating on IMDb, as well as the Audience Count on Rottentomatoes. In accordance with the legend, small, dark colored points represent a low audience count, while larger, lighter blue colored points represent a high audience count.

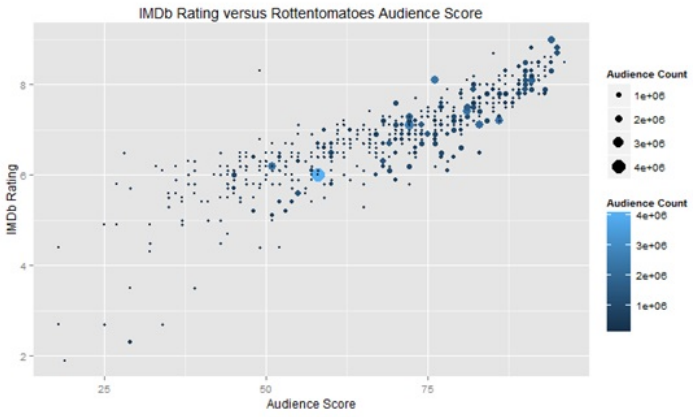
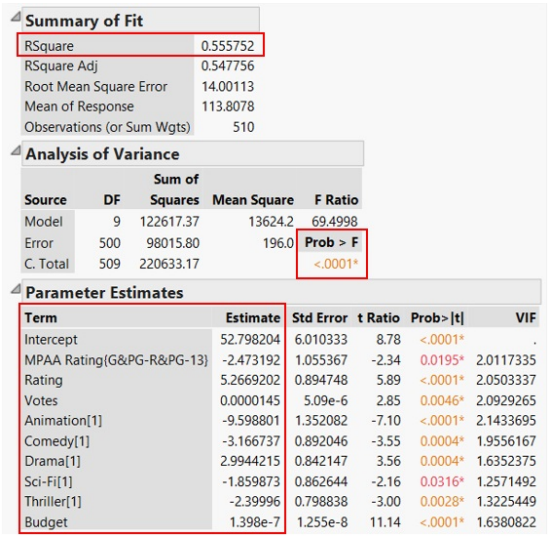


Figure 2: Removed Outliers

With now 540 observations, the average length of the movies in the data set is 113.1 minutes long and the average IMDb rating is 6.9 with the average number of IMDb votes at 210,085. The average year for these movies is 2009. The average budget for the movies, excluding the 26 movies which did not have budget information, at $81,117,386. The average All Critic Score, and average counts, are 60.65% and 180 votes. The average Top Critic Score, and average counts, are 56.8% and 38.73 votes. The average Audience Score, and average counts, are 67% and 384,745 votes. The next few statistics area counts of genre and MPAA rating to get a better idea of the number of movies with a specific genre affiliation and the number of movie with the different MPAA ratings. To reiterate, a movie can have more than one genre, so the total number of movies across all genres will not add up to 540. There are 198 Action, 188 Adventure, 70 Animation, 209 Comedy, 183 Drama, 5 Documentary, 97 Romance, 105 Sci.Fi, and 142 Thriller movies. The MPAA rating distribution is 14 G, 105 PG, 246 PG-13, 171 R, 2 Not Rated, and 2 movies without MPAA ratings. For each of the numerical variables, there did not appear to be substantial deviations from a normal, bell-shaped symmetric distribution.

Since one of the questions that was sought to address in this paper is the relationship between the length of a movie and details attributed to the movies, we fit various regression models to the data. (As mentioned above, the numerical quantities in our analysis appeared to follow an approximately normal distribution.) The first regression run was an "all-in" multiple regression. Using all variables (excluding Title) to predict the length of the movie, an R-Squared value of 0.567655 was found, and an adjusted R-Squared value of 0.548003. The R-Squared value is also known as the Coefficient of Determination, and is a measure of the percent of variation in the predicted variable that is explained by the predictor variables. In this case, 56.7655% of the variation in movie length is predicted by year, MPAA rating, votes, rating, genre, budget, All Critic Score, and All Critic Count.

The first "all-in" regression was a first step, however there is likely some multicollinearity unaccounted for. Multicollinearity is assessed in JMP through the Variance Inflation Factor (VIF) value in a regression output. A VIF score of less than 5 indicates there is likely no correlation between the specified variable and all other variables. A VIF score between 5 and 10 indicates there is likely some correlation between explanatory variables and should be investigated, although it is possible that there isn't significant multicollinearity. A VIF score greater than 10 indicates multicollinearity exists between the specified variable and one or more of the other explanatory variables. So before assessing the statistical significance of the variables, the highly correlated predictor variables must be omitted. Since Top Critic Score is a subset of the All Critic Score, it is likely these variables have some multicollinearity. Same goes for the Counts of these variables. After further investigation, it seems there is also some multicollinearity between the IMDb rating, and the Audience Score, as well as the IMDb votes and Audience Counts. The high correlation between IMDb Rating and Rottentomatoes Audience Score brings to light a new hypothesis, that Audience Score and IMDb rating are highly correlated, which would be another investigation to be covered in this research paper, but for now the "all-in" regression will remain the focus. The remaining variables do not seem correlated, because their VIF scores are less than 5, so their estimates and significance can be investigated. According to the regression output, it appears the only significant variables for predicting the length of the movie are the IMDb rating, the number of votes associated with the rating, an MPAA rating of R, the genres Animation, Comedy, and Drama, and the budget of the movie. Referring to a predictor variable as significant means at a 95% confidence level, the variable has an effect (positive or negative) on the predicted variable. This final regression has an R-Squared value of 0.560057 and the regression itself for predicting length is statistically significant. Running a quick regression analysis using only the statistically significant variables results in an R-Squared value of 0.540433. This means Rating, Votes, Animation, Comedy, Drama, and Budget predict 54.04% of the variation in the length of the movie. Which means approximately 45.96% of the variation is the length of the movie is unexplained. After some trial and error, a regression with a relatively high R-Squared value, while retaining only variables that are significant predictors of length, was produced. Predicting length with the following variables produced an R-Squared value of 0.555752: PG-13 and R MPAA ratings, IMDb rating, IMDb votes, the genres Animation, Comedy, Drama, Sci-Fi, and Thriller, and Budget. Essentially, only about 55.6% of the variation in the length of the movie is explained by the IMDb rating and votes, the genres Animation, Comedy, Drama, Sci-Fi, and Thriller, and the Budget of the movie. See Figure 3 for the JMP output of the regression.



Summary of Fit

RSquare	0.555752
RSquare Adj	0.547756
Root Mean Square Error	14.00113
Mean of Response	113.8078
Observations (or Sum Wgts)	510

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	122617.37	13624.2	69.4998
Error	500	98015.80	196.0	Prob > F
C. Total	509	220633.17		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	52.798204	6.010333	8.78	<.0001*	.
MPAA Rating(G&PG-R&PG-13)	-2.473192	1.055367	-2.34	0.0195*	2.0117335
Rating	5.2669202	0.894748	5.89	<.0001*	2.0503337
Votes	0.0000145	5.09e-6	2.85	0.0046*	2.0929265
Animation[1]	-9.598801	1.352082	-7.10	<.0001*	2.1433695
Comedy[1]	-3.166737	0.892046	-3.55	0.0004*	1.9556167
Drama[1]	2.9944215	0.842147	3.56	0.0004*	1.6352375
Sci-Fi[1]	-1.859873	0.862644	-2.16	0.0316*	1.2571492
Thriller[1]	-2.39996	0.798838	-3.00	0.0028*	1.3225449
Budget	1.398e-7	1.255e-8	11.14	<.0001*	1.6380822

Figure 3: Predicting Length

The regression Analysis of Variance shows the regression is significant with a p-value $< .0001$. All of the included variables are significant in predicting the length of the movie. This output also includes the parameter estimates for the multiple regression equation. Expanding on these estimates, a better grasp of the effect of these variables on the length of the movie can be seen. Looking at Rating first, an increase in 1.0 IMDb rating, holding everything else constant, is predicted to increase the length of the movie by 5.267 minutes. The MPAA Rating is statistically significant when grouping G and PG together and PG-13 and R together. This variable indicates that when a movie with a G or PG rating is predicted to be 2.47 minutes shorter, all else constant, while a movie with a PG-13 or R rating is predicted to be 2.47 minutes longer. Moving on to another parameter estimate, adding the Sci-Fi genre to a movie holding all else constant, which realistically mean including enough features in the movie to have it be given the Sci-Fi genre, is predicted to decrease the length of the movie by 1.85 minutes (a "0" in the Sci-Fi column means that movie is not considered a "Sci-Fi"). Looking at Animation, adding the animation genre to a movie is predicted to decrease the length of the movie by 9.6 minutes. Below is a graph of length against rating with votes and budget, which is a great representation of the above regression, given limited time and resources. Lastly, looking at Budget, an increase in \$1,000,000 US is predicted to increase the length of the movie by 0.1398 minutes (roughly 8.3 seconds). Note that not all parameter estimates were touched on, only a few examples were made to demonstrate how this output is interpreted.

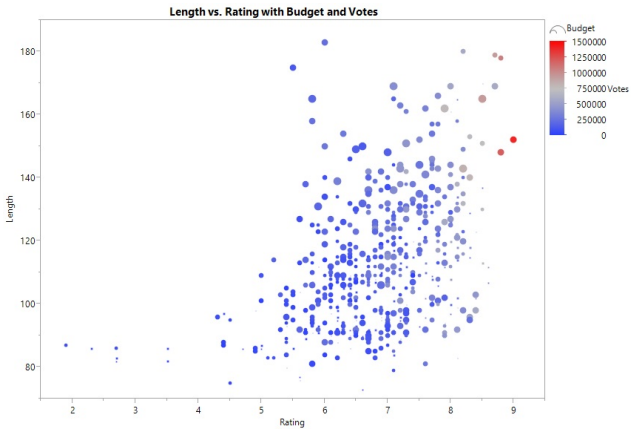
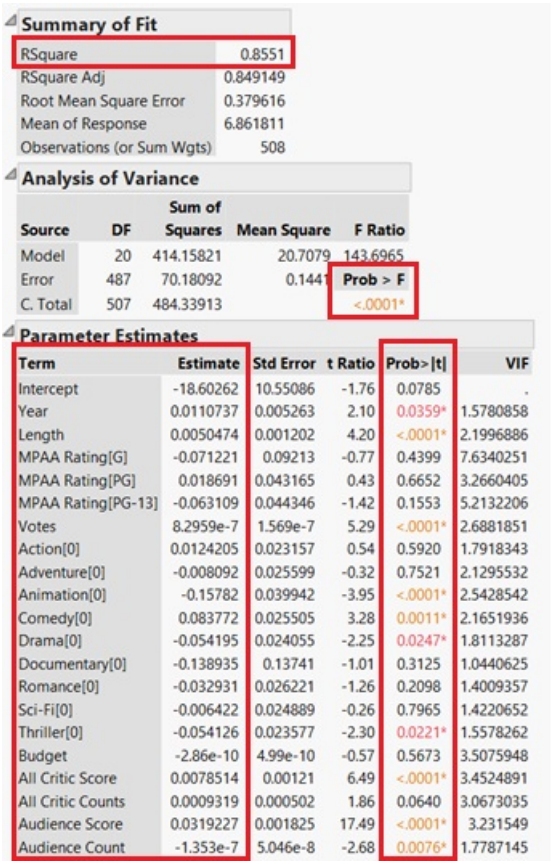


Figure 4: Length vs. Rating with Budget and Votes

Referring to Figure 4, length on the y-axis as the predicted variable and Rating on the x-axis as the predictor variable. This graph also includes Budget which determines the size of the plotted point. The larger the point the higher the budget of the associated movie. It also includes the number of IMDb votes which determines the color of the plotted points. The line is a line of best fit, fitting Rating to Length, also giving a standard error for the regression given by the grey shading about the line. Keep in mind this line of best fit only incorporates Length vs Rating, and does not include any other variables in its prediction equation. This graph is only to demonstrate the positive linear relationship Rating, Budget, and Votes have with Length. As Rating increases, Length increases, but it is also noticeable that as budget increases length increases as well. The smaller dots seem to be at the lower end of length, while the larger dots appear to be at the higher end of length. So as budget increases, length increases. Lastly, the bluer dots seem to be at the lower end of length while the grayer and redder dots tend to be at the higher end of length. So as votes increase, length increases as well.

Referring back to the multicollinearity discovered during the first regression, taking a deeper dive into using movie attributes to predict the IMDb Rating might bring interesting results. The multicollinearity between IMDb rating and Audience Score seemed interesting enough to run another regression using the "all-in" method trying to predict IMDb rating, leaving out Top Critic Score and Top Critic Count due to its multicollinearity with All Critic Score and All Critic Count. This regression resulted in a 0.8551 R-Squared value, which is remarkably high in comparison to the previous regression predicting movie length. Using the "all-in" method, approximately 85.5% of the variation of the IMDb rating can be explained by Year, Length, MPAA rating, etc. Figure 5 is the JMP output where more information about the regression is located.



Summary of Fit

RSquare	0.8551
RSquare Adj	0.849149
Root Mean Square Error	0.379616
Mean of Response	6.861811
Observations (or Sum Wgts)	508

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	20	414.15821	20.7079	143.6965
Error	487	70.18092	0.1441	Prob > F
C. Total	507	484.33913		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-18.60262	10.55086	-1.76	0.0785	.
Year	0.0110737	0.005263	2.10	0.0359*	1.5780858
Length	0.0050474	0.001202	4.20	<.0001*	2.1996886
MPAA Rating[G]	-0.071221	0.09213	-0.77	0.4399	7.6340251
MPAA Rating[PG]	0.018691	0.043165	0.43	0.6652	3.2660405
MPAA Rating[PG-13]	-0.063109	0.044346	-1.42	0.1553	5.2132206
Votes	8.2959e-7	1.569e-7	5.29	<.0001*	2.6881851
Action[0]	0.0124205	0.023157	0.54	0.5920	1.7918343
Adventure[0]	-0.008092	0.025599	-0.32	0.7521	2.1295532
Animation[0]	-0.15782	0.039942	-3.95	<.0001*	2.5428542
Comedy[0]	0.083772	0.025505	3.28	0.0011*	2.1651936
Drama[0]	-0.054195	0.024055	-2.25	0.0247*	1.8113287
Documentary[0]	-0.138935	0.13741	-1.01	0.3125	1.0440625
Romance[0]	-0.032931	0.026221	-1.26	0.2098	1.4009357
Sci-Fi[0]	-0.006422	0.024889	-0.26	0.7965	1.4220652
Thriller[0]	-0.054126	0.023577	-2.30	0.0221*	1.5578262
Budget	-2.86e-10	4.99e-10	-0.57	0.5673	3.5075948
All Critic Score	0.0078514	0.00121	6.49	<.0001*	3.4524891
All Critic Counts	0.0009319	0.000502	1.86	0.0640	3.0673035
Audience Score	0.0319227	0.001825	17.49	<.0001*	3.231549
Audience Count	-1.353e-7	5.046e-8	-2.68	0.0076*	1.7787145

Figure 5: Length vs. Rating with Budget and Votes

Beyond regression analysis, more questions can be answered with this data. In the above regression, All Critic Score and Audience Score are statistically significant in predicting the IMDb rating. Analyzing the correlation between IMDb rating and the three Rottentomatoes ratings will likely produce fascinating results. Interestingly enough, as discussed in the introduction, there are more than one correlation calculation. The traditional correlation calculation is the Pearson "r" correlation coefficient, but there is also Spearman ρ Rank correlation coefficient, Kendall τ Rank correlation coefficient, and many more. For the purpose of this research paper, the focus will be on Pearson, Spearman, and Kendall correlation calculations.

3 Correlation

Correlation, in general, is a classification that measures the strength of association between two variables. Correlation is most commonly used in linear

relationships between two variables but is not limited to linear relations. The correlation value measure varies between -1 and +1. The closer the correlation coefficient is to ± 1, the stronger the association between the two variables. As the correlation gets closer to 0, the weaker the association becomes between the two variables. A correlation with the value of -1 or +1 is said to have perfect association. For most correlation coefficient, Jacob Cohen's standards are often used to evaluate the correlation coefficient to determine the strength of the relationship, or the effect size. The standards are coefficients between ± .10 and ± .29 represent a small association; coefficients between ± .30 and ± .49 represent a medium association; and coefficients above, or below if negative correlation, .50 represent a large association or relationship. Although there are more than 20 ways to calculate correlation, typically, three types of correlations are used: Pearson correlation, Spearman correlation, and Kendall rank correlation. Determining how the correlations are calculated, the assumptions associated with these correlations, as well as when to use them is beneficial to the purposes of this paper.

4 Pearson r Correlation Coefficient

The Pearson "r" correlation is most often used in statistics in regards to linear relationships. This correlation measures the strength of the relationship between linearly related variables. For example, determining the linear relationship IMDb rating and Audience Score. Assumptions for the Pearson "r" correlation are that both variables should be normally distributed, have linearity, and homoscedasticity. Linearity is simply assuming a straight line relationship exists between the means of each of the variables. Homoscedasticity assumes the variability around the regression line is constant. One of the downsides to using the Pearson correlation coefficient is the influence of outliers. Due to the way the Pearson correlation coefficient is calculated, specifically the sum of the values and sum of the squared values, outliers can significantly influence the coefficient value. The Pearson correlation coefficient also requires some assumptions of the data that can be difficult to fulfill. However, this coefficient is extremely powerful when the objective is to determine linear relationships between variables. Below is the formula for calculating the Pearson r correlation coefficient.

$$r = \frac{n\sum(x_i y_i) - \sum(x_i)\sum(y_i)}{\sqrt{[n\sum(x_i^2) - (\sum x_i)^2][n(\sum y_i^2) - (\sum y_i)^2]}}$$

Where:
r = the Pearson r correlation coefficient
n = the number of values in each data set (these must be equal)
x_i and y_i are the two variables of interest.

The Pearson correlation is used when the relationship between the two variables of interest are predicted to be linear.

A quick example to demonstrate how the Pearson correlation coefficient is calculated. This is also to create a benchmark to compare the three correlations using the same data points. The following is the example data points and work to calculate the Pearson correlation coefficient:

x	y
0	0
1	-14
2	-11
3	-3
4	6
5	8
6	7

Below is a simple plot of the above data.

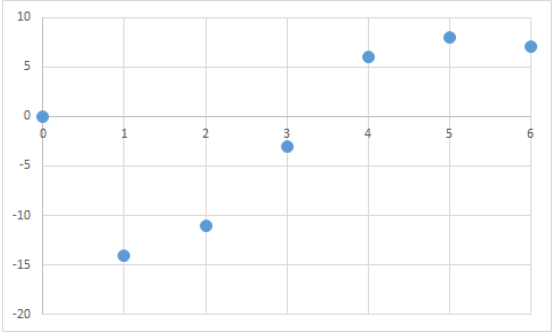


Figure 6: Toy Example Graph

The following calculations are necessary for computing the Pearson correlation coefficient:

$$\sum x = 21, \sum y = -7, \sum x^2 = 91, \sum x^2 = 475, \sum xy = 61$$

Once these have been calculated, plugging them into the Pearson equation results in an r value equal to 0.716327972.

$$r = \frac{((7 \cdot 61)(21 \cdot -7))}{\sqrt{[(7 \cdot 91 - 21^2)(7 \cdot 475(-7)^2)]}}$$

$$r = 0.716327972$$

The next correlation coefficient to be examined is the Spearman ρ rank correlation coefficient.

5 Spearman ρ Rank Correlation Coefficient

The Spearman ρ (rho) correlation is a type of rank correlation. A rank correlation is the study of relationships between rankings of different variables. In essence, Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two *ranked* variables. It was developed by Spearman, thus it is called the Spearman rank correlation. The spearman ρ rank correlation does not require any assumptions about the distribution of the data. It is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. Ordinal data is a type of data that allows for rank ordering, hence the rank correlation. This type of data allows the data to be sorted but does not have a relative degree of difference between them, e.g. the difference between rankings does not have an indication of the value of the ranked variables. In the example of movie data, a ranking based on the critic scores or ratings, the difference between rankings does not indicate the relative degree of difference between the score or rating values. The benefit of using the Spearman correlation coefficient, other than having no assumptions of the data due to the ranking of the data, is that it calculated based on a monotonic relationship . This means the Spearman correlation describes a general link between two variables, i.e. as one variables increases, so does the other, or as one variable decreases, so does the other. One would uses the Spearman correlation over the Pearson, to explain the variance over nonlinear interactions, such as parabolic relationships, exponential relationships, etc. The following formula is used to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:
ρ = Spearman rank correlation coefficient
d_i = the difference between the ranks of corresponding values x_i and y_i
n = number of values in each data set

To illustrate how Spearman ρ correlation is calculated, the example data will be used again. The first step is to rank each of the variables. Then calculated the difference between the rankings of each observation. Once the difference is calculated, squaring the difference and summing the differences obtains the desired value for calculating the Spearman correlation. Below is a table which demonstrates the above steps. Note that a "hat" is commonly placed on the Spearman correlation coefficient when calculating the correlation of a sample.

x	y	x rank	y rank	d_i	d_i^2
0	0	7	4	3	9
1	-14	6	7	-1	1
2	-11	5	6	-1	1
3	-3	4	5	-1	1
4	6	3	3	0	0
5	8	2	1	1	1
6	7	1	2	-1	1

$$\sum d_i^2 = 14$$

Plugging in this value in to the Spearman equation results in a $\hat{\rho}$ value of 0.75.

$$\hat{\rho} = 1 - \frac{6(14)}{(7)((7)^2 - 1)}$$
$$\hat{\rho} = 0.75$$

6 Kendall τ Rank Correlation Coefficient

Kendall τ (tau) correlation is another type of rank correlation. In essence, the Kendall rank correlation is a non-parametric test that is used to measure the degree of association between two *ranked* variables. It was developed by Maurice Kendall, which is why it is called the Kendall correlation. Like the Spearman correlation, the Kendall correlation does not require any assumptions of the data, again this is due to the values are ranked. Kendall, being another type of rank correlation, uses the total number of possible pairings between two data sets of the same size, n. The total number of pairings is $\frac{n(n-1)}{2}$, also known as n choose 2. Essentially, each observation is compared to every other observation in the data set. The comparisons are how the coefficient is calculated. The following formula is used to calculate the value of Kendall rank correlation:

$$\tau = \frac{\sum n_c - \sum n_d}{\frac{1}{2}n(n-1)}$$

Where:
τ = Kendall Rank correlation coefficient
n_c = total number of concordant
n_d = total number of discordant
n = number of values in each data set

Concordant means a state in which things agree and do not conflict with each other. So in terms of mathematics correlation calculations for Kendall, concordant means the observations in comparison to the rest of the observations are in chronological rank order, i.e. 1 is a higher rank than the following

rank 2 and are thus concordant. Discordant means a state in which things do not agree and do conflict with each other. So in terms of mathematics correlation calculations for Kendall, discordant means the observation in comparison to the rest of the observations are out of chronological rank order, i.e. 2 is a lower rank than the following rank 1 and are thus discordant. A more robust approach is to compare the rank orders between the variables, which is what is being done for the movie data. For example, suppose student A exercises more than student B, meaning A is higher ranked than B, in a data set. Checking whether it is also the case that student A smokes more than student B is possible. If it turns out that student A smokes more than student B, then A and B would have the same relative rank orders, and student A and student B are concordant pairs with respect to the variables Exercise and Smoke. If student A turns out to smoke less than student B, then A and B are discordant pairs.

To illustrate how Kendall τ correlation is calculated, the example data will be used one last time. The first step for calculating the coefficient is to rank both variables from largest being ranked 1st, down to the smallest being ranked last, so that all of the X values are concordant. Once both variables have been ranked, the X variable is then sorted in order of highest ranked to lowest. After that, the Y observations are compared to one another and determine the number of concordant and discordant pairs. For the example, the X observations are all concordant and are thus ignored for calculations, and only the Y observations are considered. Now the 2 is compared down the list of Y observations. The 1 is higher ranked than the 2, so that is a discordant pair. However the remaining observations are lower ranked than the 2, so they are all concordant pairs. This is then done for the remaining observations until the last observation, which does not have a comparison because there is nothing below it in the table. Below is the table of the example observations, the corresponding rankings, and the number of concordant and discordant pairs.

x	y	x rank	y rank	n_c	n_d
6	7	1	2	5	1
5	8	2	1	5	0
4	6	3	3	4	0
3	-3	4	5	2	1
2	-11	5	6	1	1
1	-14	6	7	0	1
0	0	7	4		

After all of the concordant and discordant pairs are determined, summing the number of concordant pairs and discordant pairs gives the desired value for the Kendall correlation coefficient equation. Note that a "hat" is commonly placed on the Kendall correlation coefficient when calculating the correlation of a sample.

$$\hat{\tau} = \frac{17 - 4}{\frac{1}{2}(7)(7-1)}$$

$$\hat{\tau} = 0.6190476$$

To reiterate the different calculations and emphasize the different values obtained from these different calculations, of the same data set, the results are placed below:

Pearson r coefficient $= 0.716328$
Spearman $\hat{\rho}$ coefficient $= 0.75$
Kendall $\hat{\tau}$ coefficient $= 0.6190476$

As the coefficients clearly illustrate, the Spearman is a higher value than the Pearson, and Kendall is the lowest value. This example is designed to demonstrate the similar results as the actual movie data. Correlation calculations of the movie data resulted in Spearman with the highest coefficient, followed by the Pearson coefficient, and the Kendall coefficient as the smallest.

7 Main Results

7.1 Point Estimate of Correlation: All Data

Applying these three types of correlation coefficient calculations to the movie data is the next step to understanding the correlation between the IMDb and the Rottentomatoes rating systems. Since there are three types of ratings associated with each movie for Rottentomatoes, determining the correlations between each of these three scores and the IMDb rating is likely to produce interesting and different results for each comparison. Beyond that, applying the three different types of correlation calculations will also produce interesting and different results for each of the three coefficient calculations. After computing these coefficients in RStudio, a way of visualizing this data became a concern, and became an even bigger concern later when discussing confidence intervals for the true correlation coefficient. The most efficient way to represent this data was through bar charts because of their easiness to read, and also easiness to create. Below is the actual correlation coefficients for the three types of calculations, as well as the three different comparisons made.

	Pearson	Spearman	Kendall
IMDb vs All Critic Score	0.7832294	0.7999714	0.6170782
IMDb vs Top Critic Score	0.729730	0.7468018	0.5625987
IMDb vs Audience Score	0.8711539	0.9049165	0.7548911

Figure 7: All Data All Correlations

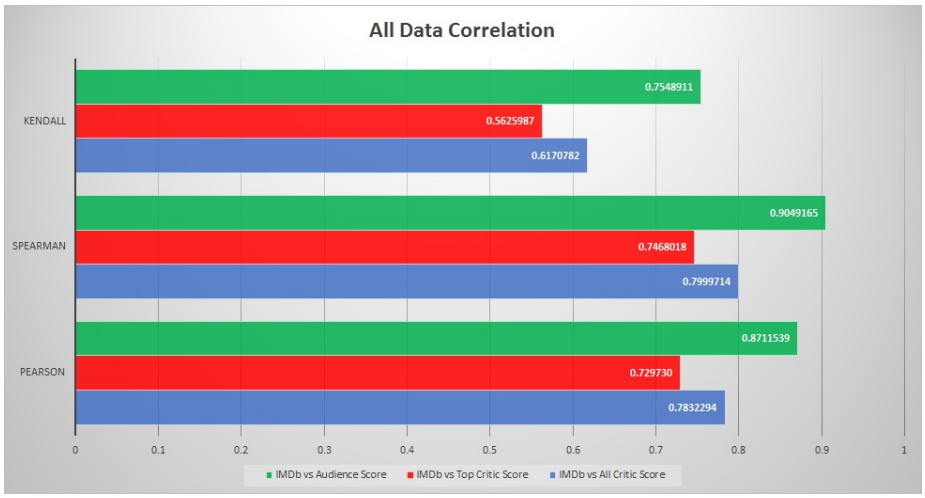


Figure 8: Bar Chart of All Data All Correlations

Observing the bar graph better illustrates the differences between the nine values (three comparisons each with three correlation calculations). It is more obvious with this graph that Kendall, overall, has the lower of the three coefficients, while Pearson trails Spearman narrowly.

The current differences between the three calculations may not be obvious. The reason for the differences between the Pearson coefficient and the Spearman and Kendall coefficients is because Pearson correlation looks for strictly linear relationships and uses the actual observation values themselves, while the Spearman and Kendall coefficients both look for monotonic relationships and rank the observations based on the observation values. Now, the main difference between Spearman and Kendall coefficients is that the Spearman correlation coefficient is calculated by determining the difference between rankings of the two variables of the data set, while the Kendall correlation coefficient is calculated by determining the concordance of the rankings. Calculating the difference between two rankings is going to result in a different value than calculating the total number of concordant and discordant pairs by comparing each pair to every other pair.

Even more information can be drawn from a plot of IMDb Rating against the three Rottentomatoes score. Below is the JMP plot.

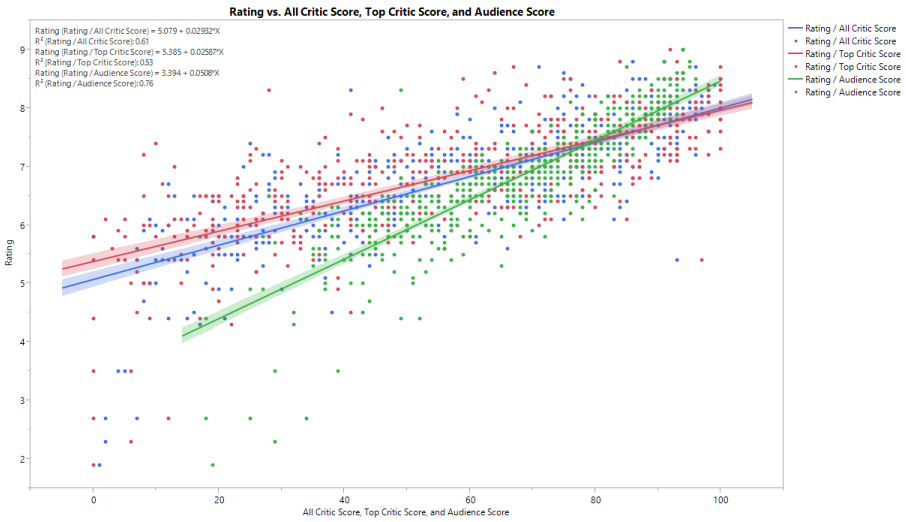


Figure 9: IMDb vs Rottentomatoes Pearson Visual

As this graph can help illustrate, the Pearson r correlation coefficient is much larger for IMDb rating versus Audience Score, as compared to All Critic Score and Top Critic Score. The green dots and line, representing the Audience Score plotted against the IMDb Rating, is a great representation of a highly correlated linear relationship. The "tightness" of the green points about the green line indicates a high correlation. The R-Squared value and the linear regression equation for this line and the other two are in the top left hand corner of Figure 9. As mentioned earlier, R-Squared is just the squared Pearson r correlation coefficient also known as the coefficient of determination. According to this value, approximately 76% of the variation in IMDb rating is explained by Audience Score, while only 61% and 53% of the variation of IMDb rating is explained by the All Critic Score and Top Critic Score, respectively. It is evident that because the red and blue points are not as "tight" about the red and blue lines, respectively, the Pearson correlation coefficient is smaller for these two variables.It is also evident in this graph that the Top Critic Score is a subset of the All Critic Score because the red and blue lines seem fairly parallel.

7.2 Point Estimate of Correlation: Subsets of the Data

After seeing the correlations between all of the data points, is it possible for the correlation between subgroups, such as MPAA ratings, groups of year, groups of length, or genre, be higher or lower? Breaking the entire data set into a few subsets at a time will help answer this question.

The first grouping will be by MPAA rating. Placing the movies into subsets based on their MPAA rating, will allow for correlation analysis based on MPAA rating. The four MPAA ratings uses are G, PG, PG-13 and R. Below is the bar graph output of the four ratings.

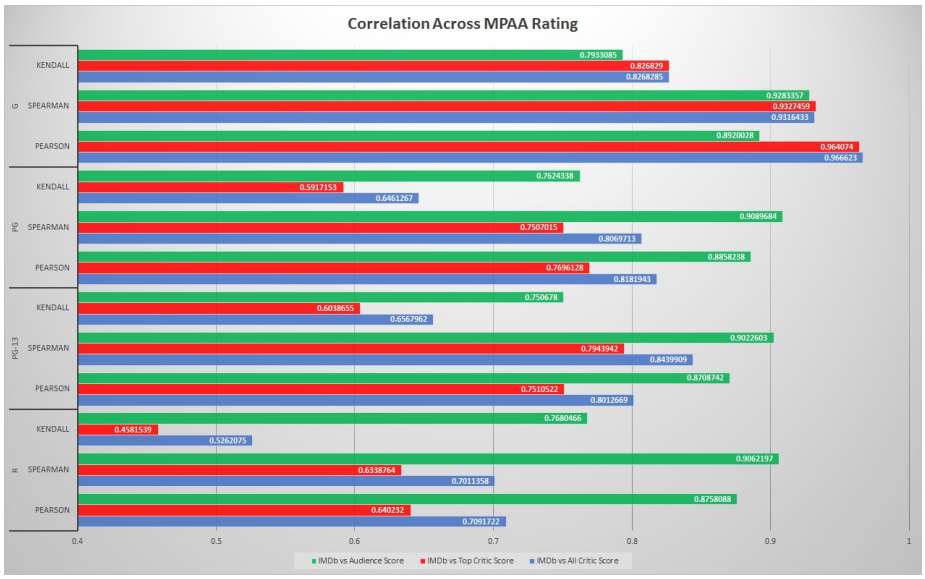


Figure 10: Bar Chart of MPAA Rating Across All Correlations

This grouping of the data, in reference to all of subsets, follows similar to trends as all of the data did. In all cases, except for G, Audience Score had a higher correlation with IMDb rating than All Critic Score and Top Critic Score did. The exception is G, which is likely due to the fact that there were only 14 movies with the MPAA rating of G in this movie data set. It is also possible that critic scores have a higher correlation with IMDb ratings naturally. However, it can be seen that for most of the coefficient calculation the Spearman coefficient is the highest, next to the Pearson coefficient, and Kendall having the smallest coefficient.

Below is the Pearson correlation visualization of each of the MPAA rating subsets.

Figure 11: IMDb vs Rottentomatoes Pearson Visual of MPAA Rating

Similar to the all data graph, this graph helps illustrate the Pearson r correlation coefficient is much larger for IMDb rating versus Audience Score, as compared to All Critic Score and Top Critic Score. Again the "tightness" of the green points about the green line indicates a high correlation, in comparison to the red and blue points and lines. The R-Squared values for this line and the other two are in the top left hand corner of Figure 11. Picking PG-13 as an example, according to the R-Squared value, approximately 76% of the variation in IMDb rating of movies rated PG-13 is explained by Audience Score, while only 64% and 56% of the variation of IMDb rating is explained by the All Critic Score and Top Critic Score, respectively. It is evident that because the red and blue points are not as "tight" about the red and blue lines, respectively, the Pearson correlation coefficient is smaller for these two variables.It is also evident in this graph that the Top Critic Score is a subset of the All Critic Score because the red and blue lines seem fairly parallel.

Potentially, this trend (in reference to Spearman being greater than Pearson and Pearson being greater than Kendall and Audience Score versus IMDb being the most correlated amoung the three correlation calculations) can exist through all of the subsets of the data, but to be sure, looking at a few more can confirm this.

Below is the grouping of movies based on movie length. The length cut offs were based on trying to get an equal amount of movies into each group. The five subsets are movies with length between 73 and 95 minutes, between 96 and

105 minutes, between 106 and 115 minutes, between 116 and 130 minutes, and between 131 and 183 minutes. Below is the bar chart of the correlations for each of the three correlation calculations and the five subsets of length.

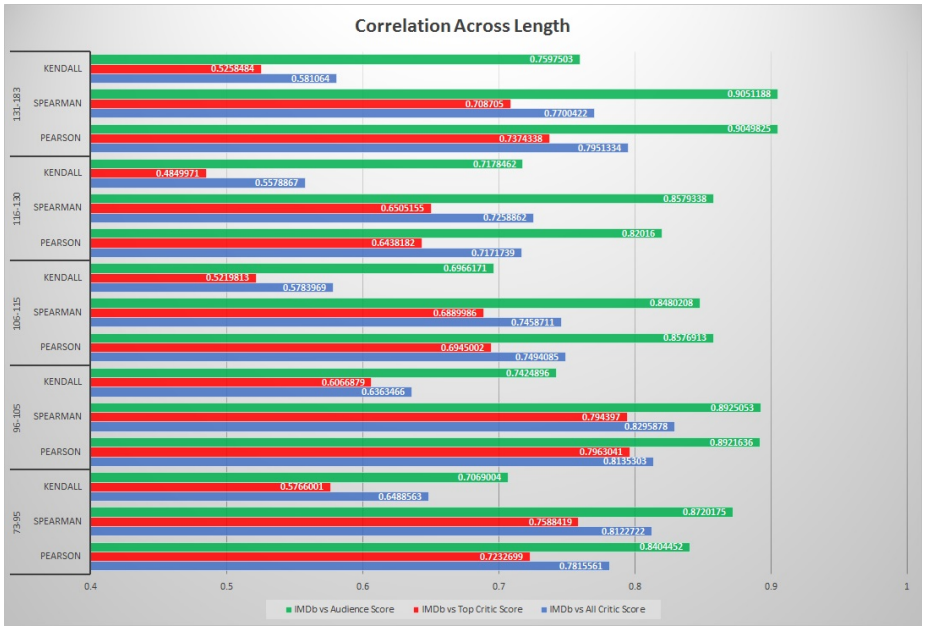


Figure 12: Bar Chart of Length Across All Correlations

Many of the same observations can be made about this grouping of the data. The Audience Score seems to be more highly correlated with IMDb rating (the green bar), across all correlation calculations, in comparison to the All Critic Score and the Top Critic Score. Also the All Critic Score has a higher correlation with IMDb rating than the Top Critic Score. Additionally, Spearman seems to have a higher value than Pearson, and Kendall has the smallest coefficient.

Plotting the points, with a line to demonstrate the Pearson correlation, will help reiterate the Pearson results from the bar graph.

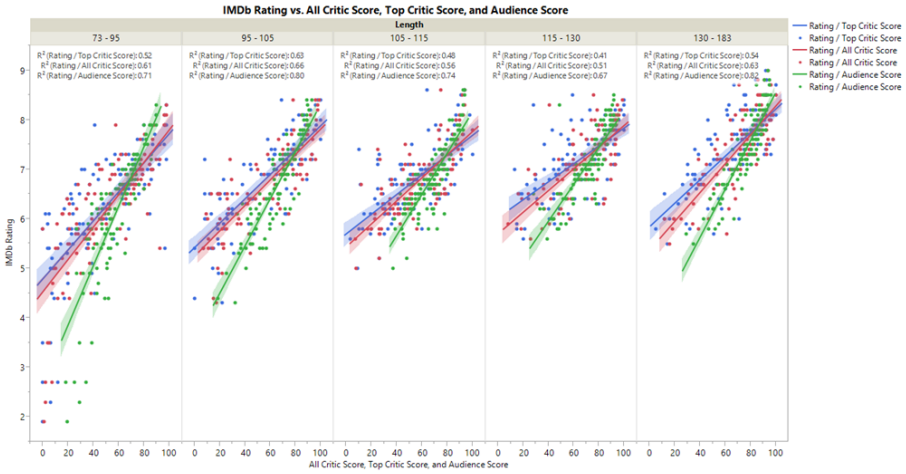


Figure 13: IMDb vs Rottentomatoes Pearson Visual of Length

This plot provides similar results as the previous Pearson plot. The green dots seem to be "tighter" to the green line in comparison to the red and blue dots to the red and blue lines.

The next grouping of the data is based on year. The year cut offs were based on trying to get an equal amount of movies into each group. The five subsets are movies with year between 1996 and 2006, between 2006 and 2009, between 2010 and 2011, between 2012 and 2013, and between 2014 and 2015 minutes. Below is the bar chart of the correlations for each of the three correlation calculations and the five subsets of year.

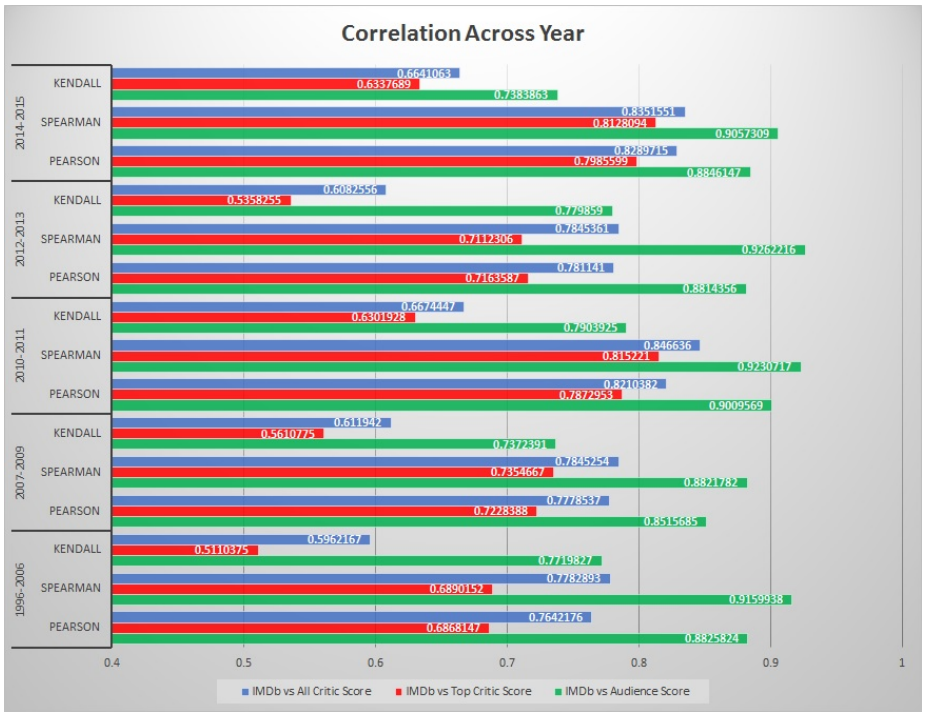


Figure 14: Bar Chart of Year Across All Correlations

Many of the same observations can be made about this grouping of the data. The Audience Score seems to be more highly correlated with IMDb rating (the green bar), across all correlation calculations, in comparison to the All Critic Score and the Top Critic Score. Also the All Critic Score has a higher correlation with IMDb rating than the Top Critic Score. Additionally, Spearman seems to have a higher value than Pearson, and Kendall has the smallest coefficient.

Plotting the points, with a line to demonstrate the Pearson correlation, will help reiterate the Pearson results from the bar graph.

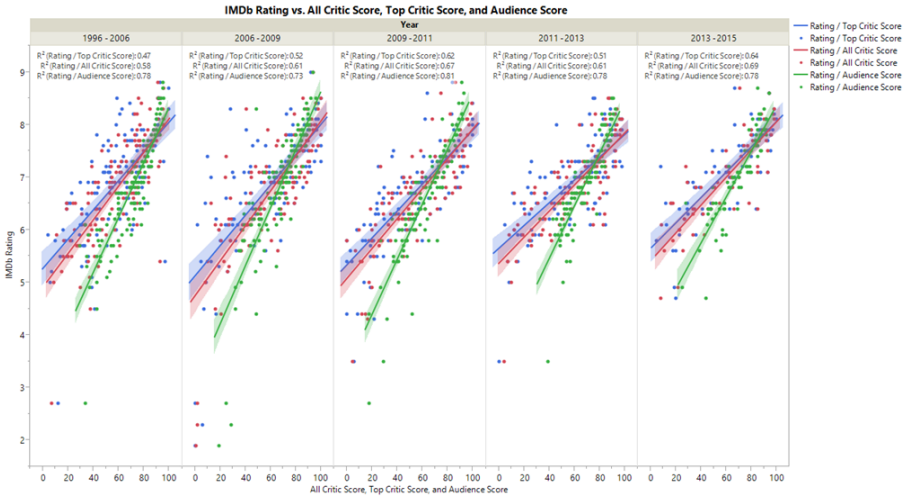


Figure 15: IMDb vs Rottentomatoes Pearson Visual of Year

Again, this plot provides similar results as the previous Pearson plot. The green dots seem to be "tighter" to the green line in comparison to the red and blue dots to the red and blue lines, respectively. The coefficient of determination for each comparison of each group is located in the top left of each of the plots.

The next grouping of the data is based on genre. Each group is determined by the genre associated with each movie. The eight subsets of movies are Action, Adventure, Animation, Comedy, Drama, Romance, Sci-Fi, and Thriller. Below is the bar chart of the correlations for each of the three correlation calculations and the eight subsets of genre.

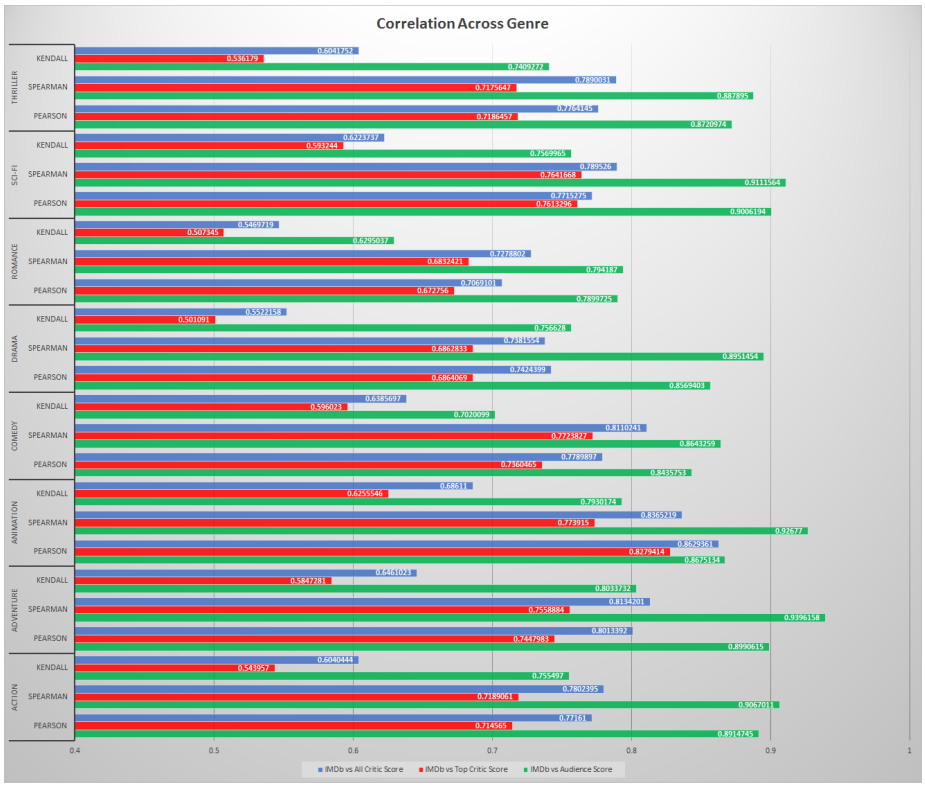


Figure 16: Bar Chart of Year Across All Correlations

Again, this grouping has many of the same observations as the previous groupings of the data. The Audience Score seems to be more highly correlated with IMDb rating (the green bar), across all correlation calculations, in comparison to the All Critic Score and the Top Critic Score. Also the All Critic Score has a higher correlation with IMDb rating than the Top Critic Score. Additionally, Spearman seems to have a higher value than Pearson, and Kendall has the smallest coefficient.

Plotting the points, with a line to demonstrate the Pearson correlation, will help reiterate the Pearson results from the bar graph. Due to the number of subsets, the plot has been broken up unto two images, however, this does not take away from the analysis.

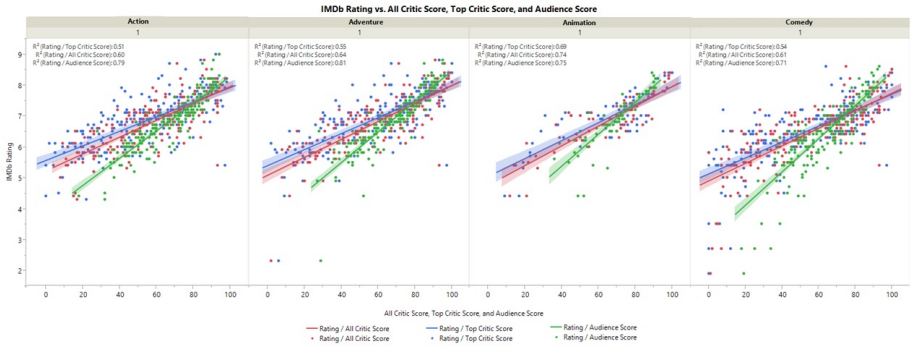


Figure 17: IMDb vs Rottentomatoes Pearson Visual of Genre pt. 1

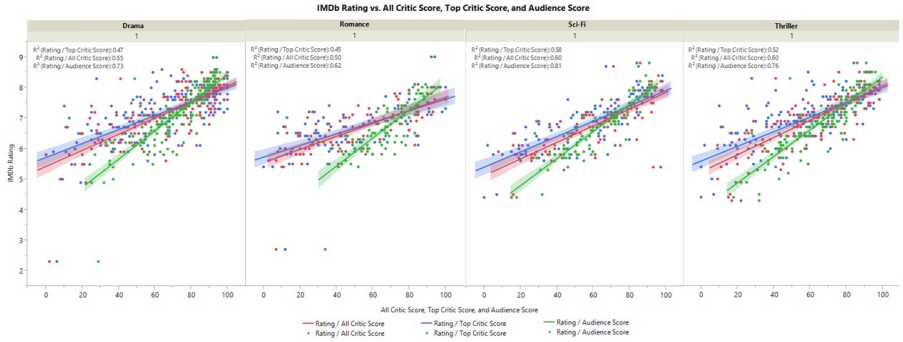


Figure 18: IMDb vs Rottentomatoes Pearson Visual of Genre pt. 2

This plot had to be manipulated in order to represent the data this way. Because of the way the variables were set up, JMP was able to plot each genre individually. All individual plots were then pasted together to create the above plots. Now, this plot, like the others, provides similar results as the all data Pearson plot. The green dots seem to be "tighter" to the green line in comparison to the red and blue dots to the red and blue lines, respectively. The coefficient of determination for each comparison of each group is located in the top left of each of the plots.

7.3 Correlation Confidence: All Data

In an attempt to go one step beyond just testing the correlation, bootstrapping was uses to approximate the standard error of the correlation coefficients to create a confidence interval for the true correlation between IMDb rating and the Rottentomatoes scores. A few steps were needed to obtain the confidence intervals. First, bootstrapping was performed to "increase" the number of samples, in order to estimate the variability of our correlation statistics. To

bootstrap, we resample our data with replacement. Each bootstrap data set was sampled with replacement of a total of the number of data points in each subset. Once the sampling was done 1000 times, the correlation of each of these samples was determined. The standard deviation of this 1000 samples was then calculated to get the approximate standard deviation of the population of movie data. A bootstrapping sample of 1000 was done for each of the three correlation calculations.

Once the standard deviation was calculated for the three correlation calculations and all of the different subsets of the data, the confidence interval could be calculated. A confidence interval of 95% was chosen because it is the most commonly used confidence level in statistics.

Because the bootstrap samples appear approximately normal, we took the critical value of 1.96 and multiplied it by the bootstrap standard deviation to estimate the margin of error for a 95% confidence interval. The margin of error is then added and subtracted from the correlation estimate to generate the 95% confidence interval for the true correlation coefficients. Again this was done for every subset tested in the previous section, as well as for each of the three types of correlation coefficients.

Below is the visual representation of the all data confidence interval.

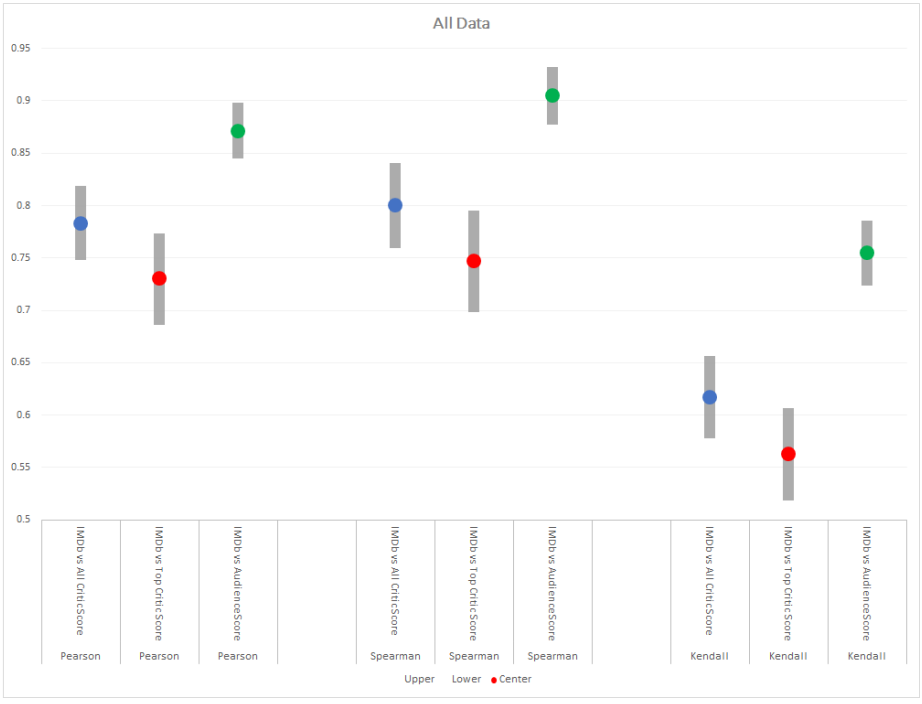


Figure 19: All Data Correlation Confidence

Based on this visualization, it is evident that Audience Score correlated with

IMDb rating is statistically different from All Critic Score and Top Critic Score correlated with IMDb rating. This graph is a great example of statistically different correlations for the different comparisons. We are 95% confident that the true population Pearson correlation coefficient between IMDb rating and Audience score is different from the true population Pearson correlation coefficient between IMDb rating and All Critic Score, as well as between IMDb rating and Top Critic Score. This is evident because the IMDb rating versus Audience Score 95% confidence interval does not overlap with either of the other two confidence intervals of the Pearson correlation calculation. This is also true for the Spearman and the Kendall correlation calculations as well. Examining subsets of the data to see if this trends consists, or if new trends emerge, will provide insight into the data itself.

7.4 Correlation Confidence: Subsets of the Data

Going beyond confidence intervals for just the entire data correlations, this analysis could benefit from exploring confidence intervals for each of the different subsets as well.

MPAA Rating visualizations are below:

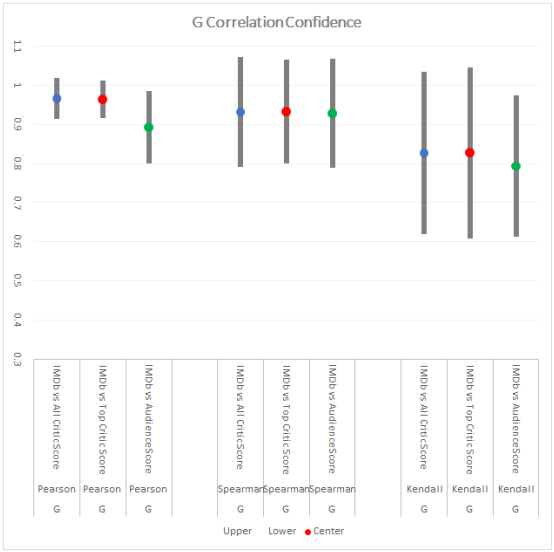


Figure 20: G: MPAA Rating Correlation Confidence

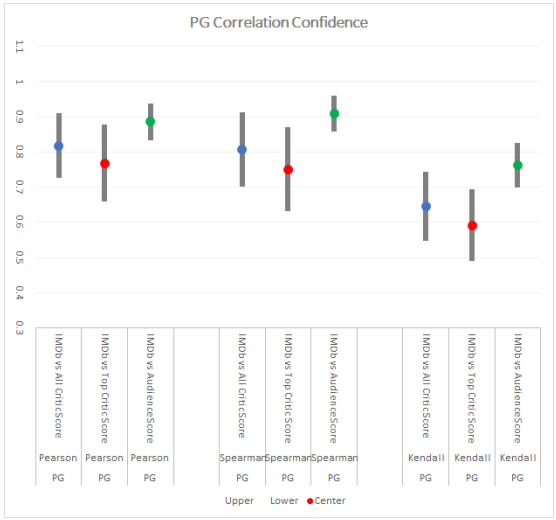


Figure 21: PG: MPAA Rating Correlation Confidence

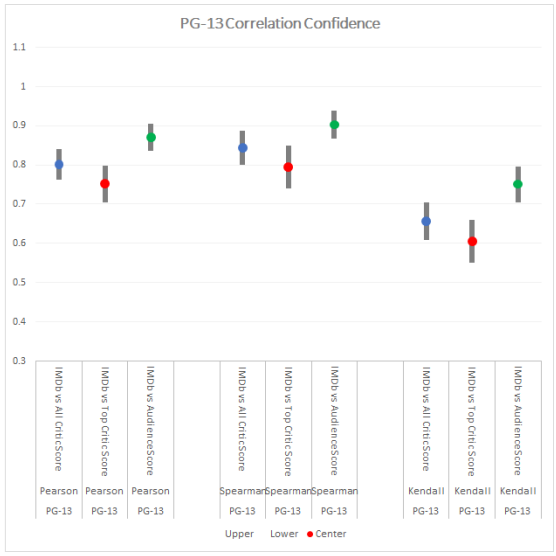


Figure 22: PG-13: MPAA Rating Correlation Confidence

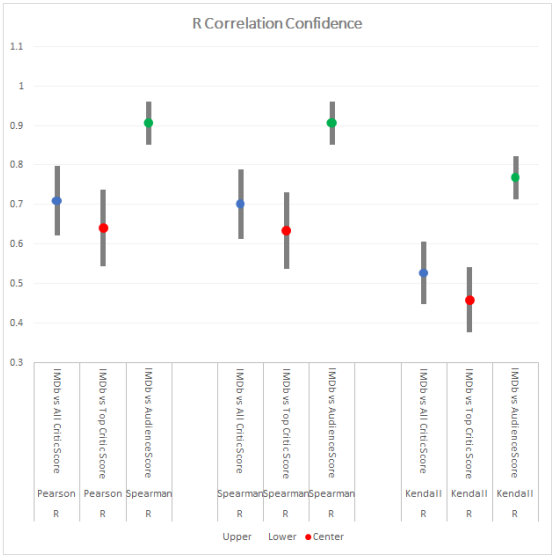


Figure 23: R: MPAA Rating Correlation Confidence

Due to the small sample size, the original point estimate of the true correlation for G rated movies is likely to vary. It is evident with the confidence intervals with both the Spearman and the Kendall coefficients of Figure 20 that there is much variability in the correlation coefficient at a small sample size. As a result the confidence interval for the Spearman and the Kendall coefficients is much wider to encompass a range of values such that there is a 95% chance that the true correlation coefficient is within the bounds.

Some of the same observations about the all data confidence intervals can be made for Figure 21. The confidence intervals for the IMDb rating versus Audience Score correlation is much narrower than that of the All Critic and Top Critic Score correlations. In this case, because the intervals overlap, no conclusions about the difference between true correlation coefficients of the three comparisons can be made.

An interesting thing to note about Figure 22: the narrow ranges of these confidence intervals indicates that at a 95% confidence level the true population correlation coefficient is relatively close to the point estimate of the correlation. In comparison, a wider confidence interval, such the Spearman or Kendall confidence interval in the subset G rated movies, the true correlation coefficient has more possible values within a 95% confidence level, so it is more likely that the population coefficient is not relatively close to the point estimate.

Figure 23 contains the first case, of a few more, in which the total data observation of statistically different correlation coefficients holds. We are 95% confident that the true population Pearson correlation coefficient between IMDb rating and Audience score is different from the true population Pearson correlation coefficient between IMDb rating and All Critic Score, as well as between

IMDb rating and Top Critic Score. Again, this is evident because the IMDb rating versus Audience Score 95% confidence interval does not overlap with either of the other two confidence intervals of the Pearson correlation calculation. This is also true for the Spearman and the Kendall correlation calculations as well. From here, we can say Audience Score is statistically higher correlated with IMDb rating for movies with the MPAA Rating of R, than All Critic Score and Top Critic Score are correlated with IMDb rating.

Similar observations can be made about the following graphs. To avoid redundancy, only key observations will be made without explanation. These explanations will be similar to the above explanations for the MPAA Rating groupings.

Length visualizations are below:

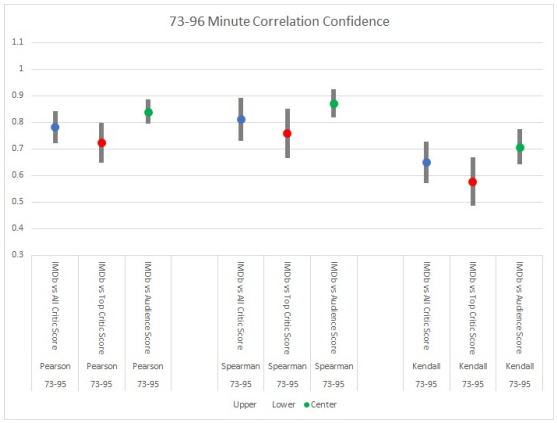


Figure 24: 73-95: Length Correlation Confidence

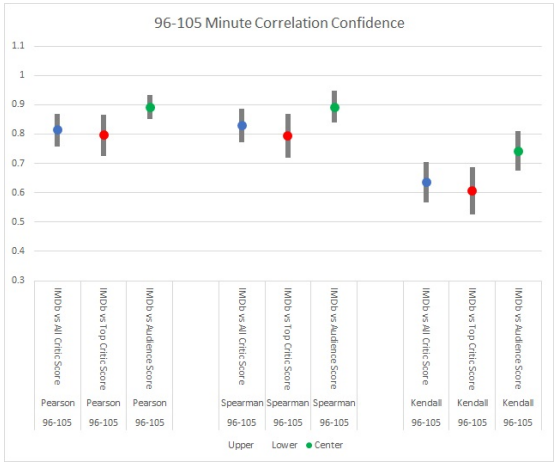


Figure 25: 96-105: Length Correlation Confidence

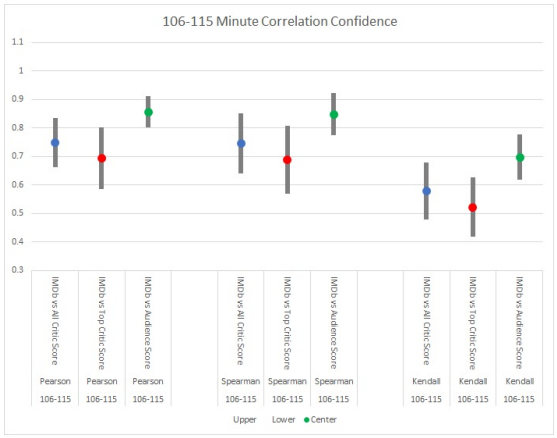


Figure 26: 106-115: Length Correlation Confidence

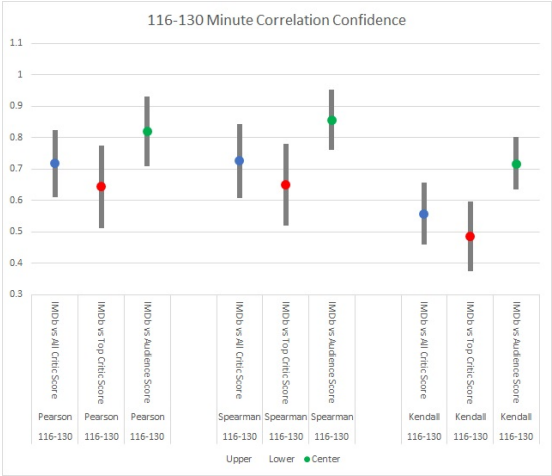


Figure 27: 116-130: Length Correlation Confidence

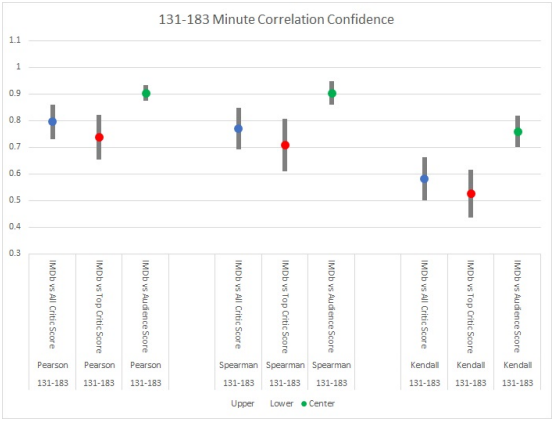


Figure 28: 131-183: Length Correlation Confidence

In all of the length groupings, Audience Score correlated with IMDb rating has the highest point estimate, as seen in the length bar chart, but is not statistically different from All Critic Score and Top Critic Score correlations.

In Figure 28, a similar situation (to movies with an MPAA rating of R) exists for movies with length between 131-183 minutes. We can say, with a 95% confidence, that Audience Score is more highly correlated with IMDb rating, for movies with length between 131-183 minutes, than All Critic Score and Top Critic Score are correlated with IMDb rating for the same length range. This holds for all three correlation calculations as well.

Year visualizations are below:

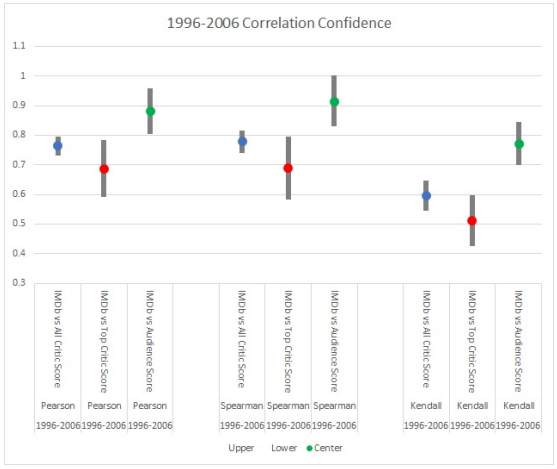


Figure 29: 1996-2006: Year Correlation Confidence

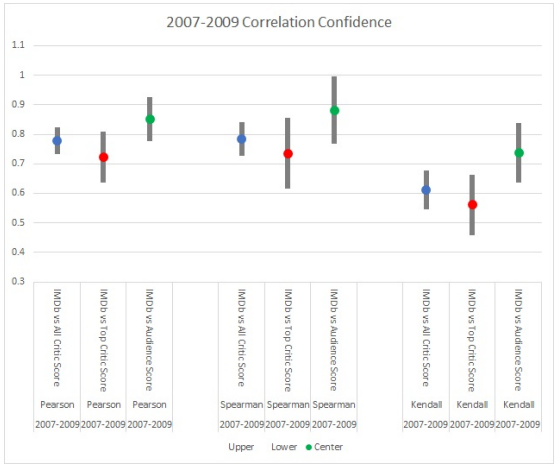


Figure 30: 2007-2009: Year Correlation Confidence

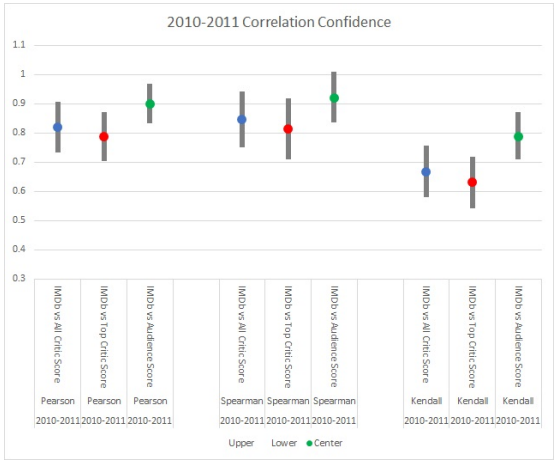


Figure 31: 2010-2011: Year Correlation Confidence

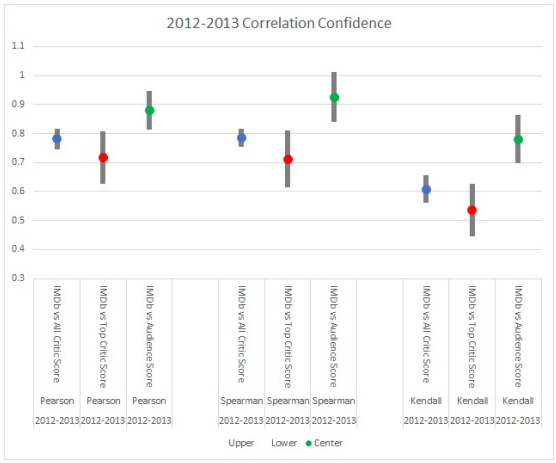


Figure 32: 2012-2013: Year Correlation Confidence

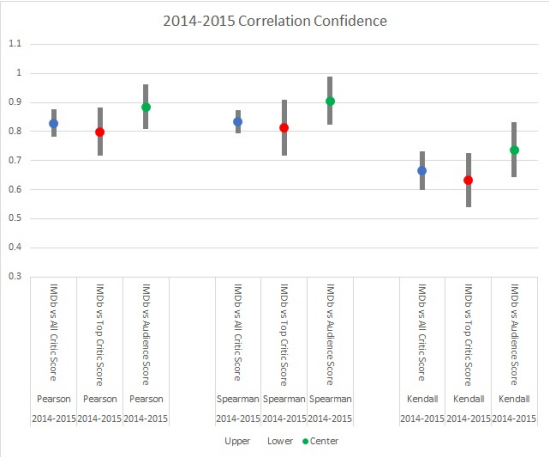


Figure 33: 2014-2015: Year Correlation Confidence

A seemingly usual occurrence for Audience Score in this subset of the data. We can say, with a 95% confidence, that Audience Score is more highly correlated with IMDb rating, for movies released between the years 1996-2006, than All Critic Score and Top Critic Score are correlated with IMDb rating for the same year range. This holds for all three correlation calculations as well.

One notable occurrence in Figure 31 and 32, is the confidence interval range for IMDb versus Audience Score under the Spearman coefficient. From the definition of correlation, the coefficient cannot go above, or below, ± 1. Because of this, the 95% confidence interval still represents an interval in which we are 95% confident that the true population correlation is within that range, except the upper bound narrows this range while retaining the 95% confidence interval. Essentially, the $+1$ limit of correlation restricts the confidence interval to a narrower range while retaining the 95% confidence.

For Figure 32, Audience Score is statistically higher correlated with IMDb rating for movies released between the years 2012 and 2013 compared to the other two Rottentomatoes scores correlated with IMDb rating.

Genre visualizations are below:

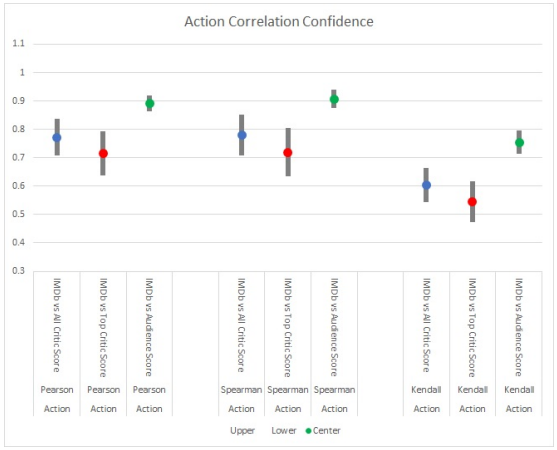


Figure 34: Action: Genre Correlation Confidence

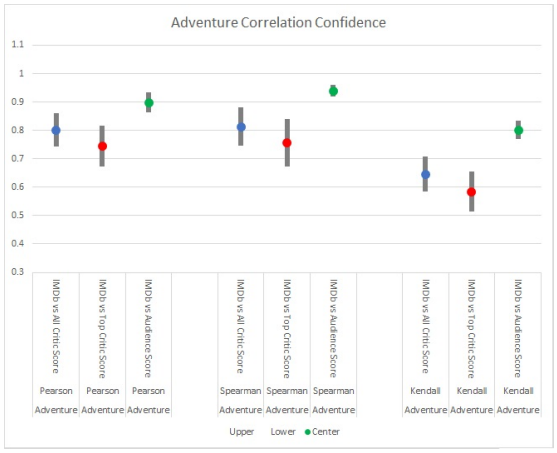


Figure 35: Adventure: Genre Correlation Confidence

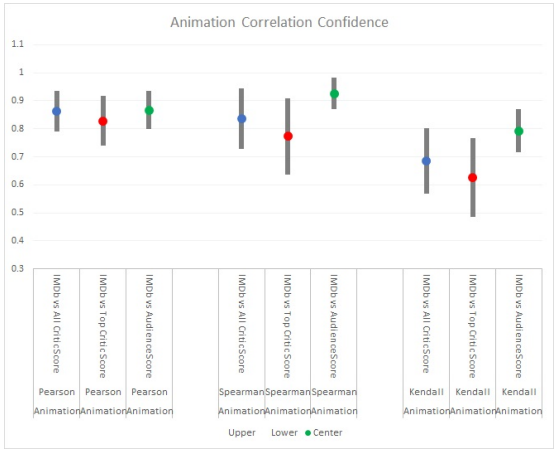


Figure 36: Animation: Genre Correlation Confidence

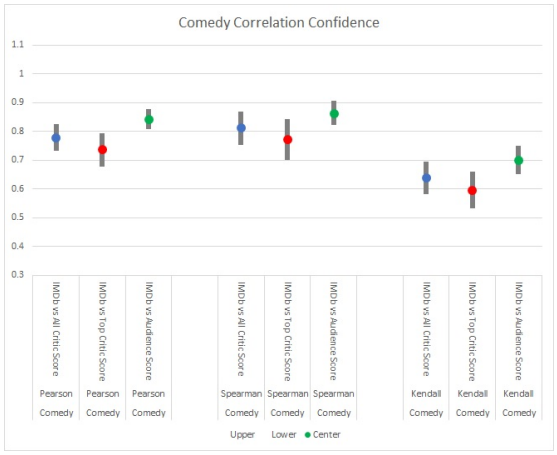


Figure 37: Comedy: Genre Correlation Confidence

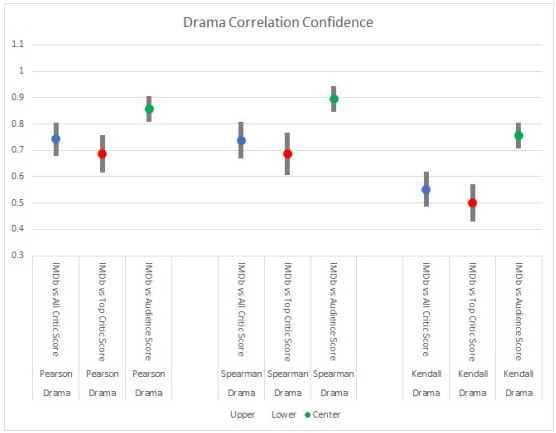


Figure 38: Drama: Genre Correlation Confidence

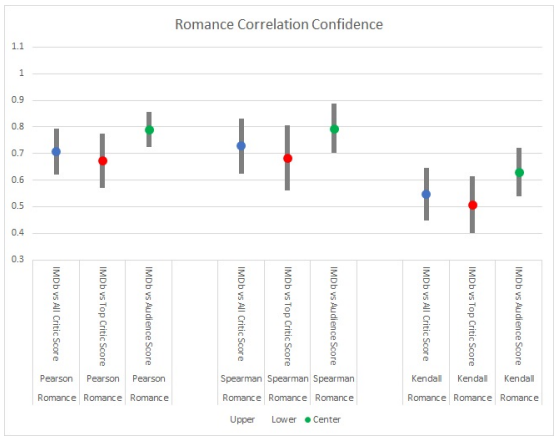


Figure 39: Romance: Genre Correlation Confidence

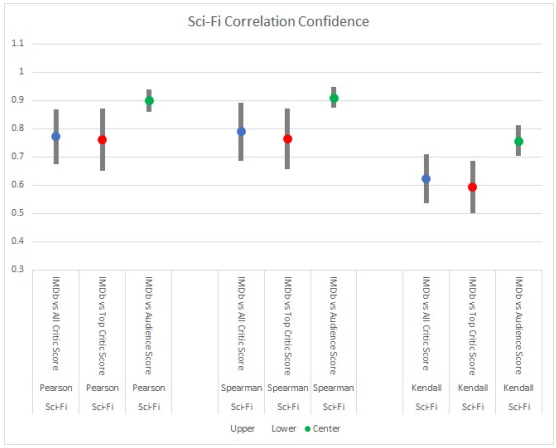


Figure 40: Sci-Fi: Genre Correlation Confidence

Figure 41: Thriller: Genre Correlation Confidence

In Figures 34, 35, and 36, many of the same observations occur. We can say, with a 95% confidence, that Audience Score is more highly correlated with IMDb rating, for movies with the Action, Animation, and Drama genres, than All Critic Score and Top Critic Score are correlated with IMDb rating for the same genre. This holds for all three correlation calculations as well.

In Figure 41, only two of the correlation calculations have Audience Score versus IMDb Rating statistically higher correlation coefficient than the other two Critic Scores correlated with IMDb rating. This holds for the Pearson calculation and the Kendall calculation, but does not hold for the Spearman correlation calculation.

8 Conclusion

Breaking down this conclusion to the many parts of analysis will help provide accurate conclusions to each of the sections of this paper.

In the prediction portion of this paper, it is evident that approximately 45% of the variation in length is left undetermined by the current data set. Perhaps this variation is better explained by director, actors, or even the entertainment distributing company (such as Lions Gate, Summit, etc.). On the other hand, approximately 15% of the variation in IMDb Rating is left unexplained. A lot of what determines the rating of the movie, based on the JMP output and the estimate p-values, is the length of the movie in minutes, the number of votes, whether or not the movie was an animation, the All Critic Score, and the Audience Score. Perhaps the remaining variation can be explained by the actors, directors, whether or not the movie was a remake, or the entertainment distribution company.

In the correlation examination portion of this paper, it is evident that the Audience Score is more highly correlated with IMDb rating, among many subsets of the data as well, than the All Critic Score and Top Critic Score are correlated with IMDb rating. It is also evident that the Spearman correlation coefficient is larger than the Pearson correlation coefficient, in most cases, which could be an indication that the relationship between IMDb rating and Rottentomatoes scores is nonlinear. It is also possible that the Spearman calculation is "seeing" a pattern within the data, one cannot, with the naked eye, easily observe. This may also be an instance of over fitting the data. However, it is more likely that then data is nonlinear as compared to the other scenarios.

In the correlation confidence portion of this paper it is evident, again, that the Audience Score is more highly correlated with IMDb rating, among many subsets of the data as well, than the All Critic Score and Top Critic Score are correlated with IMDb rating. In the overall data analysis, we saw that Audience Score was statistically more highly correlated with IMDb rating than the other two Rottentomatoes scores. It even held in some of the subsets. This portion also illustrated that the Audience correlation coefficient was often the greatest value of the three comparisons. In addition, it also reiterated that Kendall often had the lowest correlation coefficient in comparison to the other two correlation calculation methods.

9 Improvements or Further Research

Based on the limited time, resources, knowledge, and data set, some improvements and expansion of analysis can be made.

The first improvement, given more time and resources, would be to expand on the possibility of nonlinear relationships between the two rating sites. As the Spearman correlation coefficient compared to the Pearson correlation coefficient may suggest, is that the relationship between IMDb rating and the

Rottentomatoes scores may not be perfectly linear. So looking into ways to examine nonlinear relationships would be the first research topic to seek. Perhaps nonlinear relationships could provide better correlation analysis and even better predictions of a movie's length or IMDb rating.

One improvement would be collecting more movies. Increasing the observations of the data set would increase the significance of the results. A larger data set will likely lead to more accurate estimations of length and rating, as well as more accurate correlation coefficients, and even decrease the correlation confidence intervals, narrowing the search for the true population correlation coefficient.

Another improvement would be to obtain more variables for each movie. Potentially, obtaining more variables could help explain some of the unexplained variation in predicting the length or IMDb rating. There exist more than just the nine genres in this data set, so collecting more genres could increase the power of the analysis. Collecting the director, the actors and actresses, the box office totals, or the entertainment distributors could also increase the value of the analysis.

Another improvement would be statistical testing. There exists a method of statistically testing the correlation to whether or not the correlation is statistically significant. The function in RStudio exists and is the "cor.test()." This would help determine if correlations are statistically significant, and if they are statistically significant from each other correlation calculation.

Some further research would be in variable transformations. If in fact come of these variables are not normally distributed, transforming these variables to become approximately normal will definitely improve the analysis of this data. Non normal data could harm the regression analysis as well as the Pearson correlation coefficient due to the normal distribution assumption. One transformation in mind is to log transform variables. Even if they are normally distributed, log transforming the predictor variables will provide % changes in the variables. For example, if budget were to be log transformed and a regression was run with this new log(budget) variable, the resulting parameter estimate, say 'W,' would be interpreted as such: A 1% increase in budget would increase the predicted rating of the movie by $W * (100)\%$.

Another improvement for continuing research in this field is to evaluate the accuracy of the regression analysis. Other methods for predicting an output exist, such as regression trees, neural networks, etc. These would all be viable options for expanding the regression analysis for predicting the length or IMDb rating of a movie.

In general, more research options are available. Further research on other movie rating sites, or sources could be done. Perhaps other sources are more highly correlated with IMDb than Rottentomatoes was, over even more reliable. There are other valuable options for calculating the relationship between two data set, so those analyses could be performed and compared. More research on estimating the standard error of the correlation, such as using quantiling the bootstrapped samples to create the confidence intervals.

One last expansion of the research could me expanding into TV shows. Tele-

vision shows have a great presence on IMDb, so much so that not only do the television series as a whole has a rating, but individual episodes have ratings as well. This was not touched on at all during the analysis. However, interesting results could arrise from predicting the rating of television shows.

10 Appendix A

movieScore function:

```
1. movieScore == function(name)
2. {
3. name == gsub(" ", "_", name)
4. name == gsub("\\ .", "", name)
5. name == gsub(":", "", name)
6. name == gsub("-", "_", name)
7. name == gsub(",", "", name)
8. name == gsub("& ", "and", name)
9. name == gsub("The_", "", name)
10. url == paste0("http://www.rottentomatoes.com/m/",name)
11. if (url.exists(url) == TRUE) {
12. page == read_html(url)
13. page.span == xml_find_all(page,".//span")
14. page.div == xml_ind_all(page,".//div")
15. index.a == which(xml_attr(page.div, "class") == "col-full-xs visible-xs clearfix")
16. index = which(xml_attr(page.span,"itemprop") == "ratingValue")
17. table == data.frame("","","","","","","")
18. names(table) = c("title", "all_critic_score", "all_critic_count", "top_critic_score",
                     "top_critic_count", "audience_score",
                     "audience_count")
19. if (length(index) == 3)  {
20. table$title = name
21. table$all_critic_score = page %>%
22. html_nodes("span") %>%
23. .[index[1]] %>%
24. html_text() %>%
25. as.numeric()
26. table$all_critic_count = page %>%
27. html_nodes("span") %>%
28. .[index[1]+3] %>%
29. html_text() %>%
30. as.numeric()
31. table$top_critic_score = page %>%
32. html_nodes("span") %>%
33. .[index[2]] %>%
34. html_text() %>%
35. as.numeric()
36. table$top_critic_count = page %>%
37. html_nodes("span") %>%
38. .[index[2]+3] %>%
```

```
39. html_text() %>%
40. as.numeric()
41. audience.score = page %>%
42. html_nodes("span") %>%
43. .[index[3]] %>%
44. html_text()
45. audience.score = as.numeric(gsub("%","", audience.score))
46. table$audience_score = audience.score
47. audience.count = page %>%
48. html_nodes("div") %>%
49. .[index.a[length(index.a)]-1] %>%
50. html_text()
51. audience.count = gsub("User Ratings:","",audience.count)
52. audience.count = as.numeric(gsub(",","",audience.count))
53. table$audience_count = audience.count
54. }
55. else  {
56. table == data.frame(title = name, all_critic_score=NA, all_critic_count=NA,
                top_critic_score=NA, top_critic_count=NA, audience_score=NA,
                audience_count=NA)
57. }
58. }
59. else {
60. table == data.frame(title = name, all_critic_score=NA, all_critic_count=NA,
                top_critic_score=NA, top_critic_count=NA, audience_score=NA,
                audience_count=NA)
61. }
62. return(table)
63. }
```

The first line names the function. Line 2 opens the function so that the next consecutive lines of code will make up the function. Lines 3 through 9 transforms the inputted name of the movie into the form required for a URL web address. For example, if the input was "Transformers: Revenge of the Fallen", these lines would manipulate the input so that it would then be "Transformers_Revenge_of_the_Fallen". This is then pasted at the end of "www.rottentomatoes.com/m/" in line 10 completing the Rottentomatoes URL. The "/m/" is the address used for all movies on Rottentomatoes. Line 11 is the first of two "IF statements." This first if statement checks to make sure the URL we just created exists before continuing with the next lines of code. If the URL does not exist, then the function jumps to line 59 through 61, and outputs a table with the name of the movie, in the transformed URL form, in the first column and "NA"s in all 6 other columns. If the URL does exist, function continues.

In line 12, a function called "read_html" strips the graphs of the webpage, which is what we normally see when visiting a web site, down to the text that is used to create the graphics. Lines 13 and 14 then search the html code for specific values and creates a list of sections of html code with the value we were searching for. The next two lines then search these lists previously created for

different specified values and creates an "index" of the row numbers in which the values we were looking for are located. As it turns out for "Transformers: Revenge of the Fallen", the URL exists and was able to find the "span" and "div" values. The first index locating the "itemprop" value equal to "ratingValue" is located in rows 200, 209, and 217. Where the second index located the "class" value equal to "col-full-xs visible-xs clearfix" is located in row 234. These two indices are crucial for locating the scores and counts from webpage to webpage, because the location within the html code of the scores and counts changes day to day. This is caused by the variation in the "IN THEATERS," "DVD & STREAMING," and "TV SHOWS" content on every Rottentomatoes movie webpage.

Lines 17 and 18 create a table, or data frame, for outputting the scores and counts. It creates a table with 7 columns labeled Title, All Critic Score, All Critic Count, Top Critic Score, Top Critic Count, Audience Score, and Audience Count respectively. The next lines are the meat and potatoes of the function, where the desired scores and counts are obtained. Lets start with the first chunk from lines 21 to 25, and use this as our base of understanding the next chunks of code. Line 21 starts by assigning the scraped value, which is obtained in the next few lines, to the correct column in the table created in line 17. Within the html code, line 22 locates all of the "span" values, similar to what we did in line 12, then goes to the first value in the index in line 23. The "span" value in the row determined by the index is then designated as html text, which is then converted into a number. To reiterate, this number is then assigned to the All Critic Score column in the table. For our previous Transformers example, the function would located the 200th "span" value and grab the All Critic Score associated with that movie, which is 19, and assign that value in the table under All Critic Score. The function then continues to the next chunk, beginning on line 26, which then repeats the above steps but adds three to the first number in the index, because the count is always located three rows after the score, to then grab the count associated that that score. The next two chunks of code are very similar to the first two, but instead uses the second number in the index to locate the Top Critic Score and Top Critic Count and assigns them to the corresponding columns in the table.

The next chunk, beginning on line 41, works similarly to the first few, but grabs a little more than just the score. This chunk grabs the Audience Score, but we have to manipulate the value before using "as.numeric." The function progresses the same as the other chunks; it locates the "span" row of the third indexed number, then pulls the value and reads it as html text. The score, for an unknown reason, has a "%" along with the number that must be removed before assigning the score as a number and placed in the table. The function for Transformers: Revenge of the Fallen would pull out the 217th "span" value which is "58%." Then the function removes the "%" to get just "58" and then assigns the number to the corresponding column in the table. The removal of the "%" and the conversion of the text to a number is done on line 45.

The last chunk of the "meat and potatoes" is the trickiest of all. In order to locate the Audience Count, the second index, "index.a," is used. Because often

times there are more than one "div "class" value equal to "col-full-xs visible-xs clearfix, the second index has more than one number corresponding to the row in which "col-full-xs visible-xs clearfix" is located. As it turns out, the Audience Count was always the second last of these numbers. Beginning on line 47, the function looks through the html code for "div," then locates the row of the second last indexed number in "index.a." The function reads this value as html text. The value, similar to the Audience Score, had more than just the count, so we needed to remove all the extra text to get just the number before converting it to a number. For an unknown reason, the Audience Count had "User Ratings:" and the count value with commas. So in line 51, the "User Ratings:" was removed, and in line 52 the commas were removed, then the remaining value was converted to a number and assigned to the corresponding column in the table.

The last few pieces of this function are to close out the "IF statements" and to output the table. Line 54 closes the "IF statement" checking the first index for length equal to three. So if the index was not equal to three, then the function would skip down to line 55, which begins the "else" portion of the "IF statement." If the index did not have three values then the function would assign the name of the movie to the "title" column of the table, and assign "NA"s to the remaining 6 columns. Line 57 then closes the "else" portion of the second "IF statement" where line 58 closes the "if" portion of first "IF statement." Line 59 then begins the "else" portion of the first "IF statement." From here, if the URL does not exist, the function assigns the name of the movie to the "title" column of the table, and assigns "NA"s to the remaining 6 columns. Line 61 closes the "else" portion of the first "IF statement." Line 62 tells the function what exactly to output when the function is used. It output the table created through the process. Line 63 closes the entire function.

Overall, the function takes a movie title, and outputs the movie title, All Critic Score, All Critic Count, Top Critic Score, Top Critic Count, Audience Score, and Audience Count.

Once the function was up and running, a "for loop" was designed to process all of the movies in the data set. The loop ran all of the movies in the data set through the function, collected the individual outputted tables from each movie, then output that collection into an excel file to be combined with the IMDb data set. Once all of the data had been exported to excel, the movies with "NA" meant that the Rottentomatoes address was incorrect, often due to movies with the same name so the year the movie was released in theaters was added to the movie title for the web address or the address started with a reference number and then the title of the movie. These "NA" adjustments were made manually.

References

[1] Cohen, J. "Statistical power analysis for the behavioral sciences", 2nd edition (1988). Hillsdale, NJ: Lawrence Erlbaum Associates.

[2] "Correlation (Pearson, Kendall, Spearman)", *Statistics Solutions* (2012) 1-5

[3] Kendall, M.G., A New Measure of Rank Correlation, Biometrika, 30,(1938)81-93.

[4] "Kendall Rank Coefficient" *R Tutorial* (2013), 1.

[5] Pam MS, NCSP, "What is MONOTONIC RELATIONSHIP?" *Psychology Dictionary* 1

[6] Radford, Ben, "ROTTEN TOMATOES DATA IN R", *Data Anlysis* (2014), 1-2.

[7] Spearman, C. (1904), The Proof and Measurement of Association Between Two Things, American Journal of Psychology, 15, 72-101.

[8] Stephanie, "Correlation Coefficients: Find Pearsons Correlation Coefficient" *Statistics How To* (2009) 1-6

[9] http://had.co.nz/data/movies/